J Phys 80 (10), 969-974 (2006)

PHENIX, a high throughput structure determination tool for single-wavelength anomalous diffraction data

S Selvanayagam,[1] D Velmurugan,[1]* T Yamane[2] and A Suzuki[2]

[1]Department of Crystallography and Biophysics, University of Madras, Guindy Campus, Chennai - 600 025, India

[2]Department of Biotechnology and Biomaterial Science, Graduate School of Engineering, Nagoya University, Furo-Cho, Chikusa-Ku, Nagoya 464-8603, Japan

E-mail: d_velu@yahoo.com

Received 28 April 2006, accepted 28 August 2006

In Structural Genomics, one is interested in determining the structure in the fastest way to understand new folds and this has opened Throughput Crystallography'. This High Throughput structure determination requires automation to reduce the obstacles related to ation. PHENIX is a new software package for the automation of crystallographic structure solution of macromolecules. The algorithms ows one to proceed from reduced intensity data to a refined molecular model and to facilitate structure solution for both the expert and rallographers. In cases where PHENIX builds an incomplete model of less residues with or without side chains, ARP/wARP can be made ed further with this as input. Attempts are here made in extending the use of PHENIX for the high throughput structure elucidation of se of approximately 44 kDa molecular weight with lab source CuKα and CrKα anomalous scattering data sets corresponding to 1.7 and ns respectively. One manganese position initially located by PHENIX is enough to build the entire structure using the CuKα data whereas heavy is (11 heavy atoms) are required in case of CrKα data to build the entire structure.

ds: SAD, PHENIX, gluco

CS Nos.: 61.10.Nz, 87.14.Ee

Introduction

romolecular crystallography has undergone tremendous ress in the last decade. The database of known sequences owing at an exponential rate and has already resulted in plete genomic sequences for several organisms. The ional significance of coding regions of the genomes can ed when the sequence information can be interpreted ght of known structures. But only a fraction of the ed structural information is available at present. Due to reasing rate of protein structure determination, the ted efforts of the sequencing projects have been aken. But even then there remains a huge sequence data which three dimensional structural information is unknown.

The technologies needed for the steps of structure mination are becoming reliable and powerful enough that they can be linked together into an automatic sequence that can yield a structure in an automatic or nearly automatic fashion. The high-throughput structure determination will require automation to reduce the obstacles related to human intervention. Several projects are underway worldwide to automate the process of sample preparation, crystallization and data acquisition. The fully automated structure determination has supplied much of the impetus for the vision of large-scale structure determination in structural genomics. In future, macromolecular crystallography would have matured considerably to such an extent that most protein crystal structures will be solved in an almost completely automatic way. This is mostly important for researchers in pharmaceutical companies.

With recent progress in data collection techniques and the current trend towards high-throughput structure determination, the single-wavelength anomalous diffraction (SAD) approach

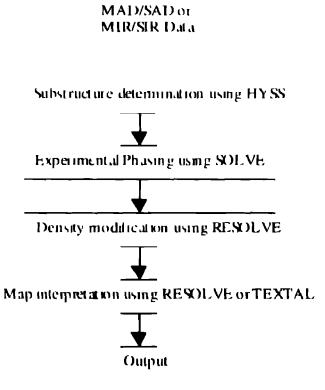
*Corresponding Author

has acquired increasing popularity and favorably competes with multiple wavelength anomalous diffraction (MAD) SAD may need more accurate intensity estimations than MAD, especially when weak anomalous scatterers such as phosphorous or sulfur are used [1,2] but does not require precise wavelength tuning and can be performed in home laboratories with Cu or Cr X-ray sources The introduction of the chromium rotating anode and designs of new synchrotron beamlines especially for the use of long-wavelength radiation are deliberately intended for the application of SAD phasing [3]

PHENIX (Python-based Hierarchical Environment for Integrated Xtallography) is a software package developed for the automatic crystal structure determination This provides the necessary algorithms to proceed from reduced intensity data to a refined molecular model [4,5]

2. Description of PHENIX program

PHENIX uses Python as the scripting language and C++ as the compiled language High-level algorithms such as complex refinement protocols or phasing procedures can be developed in a scripting language and the computation of structure factors or discrete Fourier transforms must be implemented in a compiled language for performance reasons



The algorithms implemented in PHENIX for automatic structure determination are shown in the flowchart For a given data, Hybrid Substructure Search (HYSS) makes use of Patterson and direct methods to locate anomalous scatterers or heavy atoms for experimental phasing Once the substructure has been determined, SOLVE program enables rapid configuration of jobs for experimental phasing through interfaces Initial phases can also be obtained using molecular replacement, incorporating maximum likelihood targets This can increase the success rate of this method using search models of lower structural similarity

Maximum likelihood density modification algorithms implemented in RESOLVE produce minimally biased electron density maps using either the phases obtained from experimental phasing or by molecular replacement These maps are then automatically interpreted using template matching implemented in RESOLVE and pattern recognition methods implemented in TEXTAL

The strategy concept in PHENIX provides a way to construct complex networks of tasks to perform a higher-level function For example, in SAD case, strategy concept breaks down the steps required to go from initial data to a first electron density map in various tasks The PHENIX graphical user interface (GUI) permits strategies to be visualized and manipulated The PHENIX Data Storage (PDS) in PHENIX is a data management system that oversees the information generated for each structure determination and contains a complete history of each structure solution along with all of the generated structural information PHENIX GUI also provides PYMOL, a molecular graphics system written in C and Python, allowing easy viewing of structures and maps *via* close integration with PHENIX

3. Materials and methods

As detailed papers on the success of SAD phasing using Cu radiation or synchrotron radiation with wavelength range 1 1 7Å or CrKα radiation have already appeared in the literature [6-11], this paper mainly focuses on the SAD application of PHENIX to an enzyme glucose isomerase (approximately 44 kDa molecular weight) using lab source (both CuKα and CrKα) anomalous scattering data corresponding to 1 7

Table 1 Crystallographic data

For CuKα data.	
a (Å)	93 021
b (Å)	97 948
c (Å)	102 641
Space group	I222
Resolution range (Å)	10 1 7 (1 756 1 7
Completeness (%)	99 91 (99 94)
$I/\sigma(I)$	28 38 (6 98)
Anomalous signal-to-noise ratio	0 99
For CrKα data	
a (Å)	92 872
b (Å)	97 905
c (Å)	102 710
Space group	I222
Resolution range (Å)	10-2 3 (2 377-2 3
Completeness (%)	98 47 (99 06)
$I/\sigma(I)$	58 3 (42 22)
Anomalous signal-to-noise ratio	1 39

ations This enzyme contains 388 amino acids and two
il sites, one occupied by Mn^{2+} ion and the other by Mg^{2+}
e two data sets were collected at High Intensity X-ray
ators Nagoya University using Rigaku *R*-axis IV and VII
ector system Table 1 shows the crystallographic details
se two data sets

cases where PHENIX did not build a complete model but
partial one, this model could be fed into the automatic
el building or side chain tracing by the program ARP/wARP
followed by the refinement program REFMAC [13] All the
putations mentioned here were carried out using Pentium
t

ucose isomerase, CuKα data

a manganese atom located from SAD data

NIX was run to find substructure and then for phasing and
building with the inputs of scalepack format for this data
mple letter sequence file and substructure present in this
as manganese (Mn) The imaginary component of the
nomalous scattering (f'') of manganese at this wavelength is
lectron units Initially, HYSS algorithm in PHENIX located
manganese position (along with other heavy atoms) and
his was done using SOLVE algorithm with the manganese
position After density modification and map interpretation
with SOLVE algorithm in PHENIX, the program finally built
residues in 21 chains out of a total of 388 residues Out of
1 residues, 31 residues were built with side chains At
e PHENIX gave the best score of 25 8, Figure of Merit
as 0 17, R and R_f as 48% and 50%, respectively The
ull model map correlation coefficient for this output was
0 4 A map was calculated using the SOLVE output phases and
492 peaks were above 3σ cut-off

This model was then fed to ARP/wARP [12] for automatic model building using the option 'model building using the existing model' After 50 cycles of auto-building, it was able to build 383 residues with side chains (out of a total of 388 residues) and has located 351 water atoms At this stage, the R_w and R_f values were 17 1% and 20 6%, respectively Without the water atoms R_w and R_f values were 23 3% and 24 7%, respectively The map indicated the difference densities of the missing regions and the remaining residues were modeled into this After the manual model building, the water atoms were checked and included if necessary and 25 cycles of maximum-likelihood refinement were performed using REFMAC [13] The final R_w and R_f values were 17 3 and 19 4%, respectively Table 2a details the PHENIX and ARP/WARP results

(ii) Using 11heavy atoms

Using the same data, PHENIX was run with the substructure containing eleven heavy atoms (9 sulphurs in Cys and Met residues + 1Mn + 1Mg) The imaginary component of the anomalous scattering (f'') of sulphur at this wavelength is 0 57 electron units After the substructure finding, phasing and model building, PHENIX built 341 residues in one chain out of 388 residues Out of these 341 residues, 335 residues were with side chains The overall model map correlation coefficient for this output was 0 71 A map was calculated using the SOLVE output phases and 557 peaks were above 3 σ cut-off This model was then fed into ARP/wARP which built 382 residues with side chains and located 354 water atoms Manual model building was carried out for the missing residues and the water atoms

Table 2a Details of PHENIX and ARP/wARP glucose isomerase CuKα data, one Mn position, R_w and R_f are in %

PROGRAM	Resolution limit		10-1 7 A		
PHENIX	One peak		Score = 25 8 FOM = 0 17		SOLVE MCC = 0 2990 492 peaks
	RESOLVE built		191 (31 with side chains)	R_w = 48 R_f = 50	Overall model MCC = 0 4
		Initial	R_w = 47 6	R_f = 46 7	
	No of auto building cycles		10		
	No of REFMAC cycles in each auto building cycle		5		
ARP/wARP		Final	R_w = 17 1	R_f = 20 6	
	Connectivity Index		0 99		
	No Chains		2		
	No Res Built		383 (with side chains)		
	Water atoms		351		
	R_w and R_f without water atoms		**R_w = 23.3**	**R_f = 24 7**	
Final model with solvent atoms			**R_w = 17.3**	**R_f = 19 4**	
			r m s deviation of backbone atoms(1OAD) 0 150 Å (1MNZ) 0 269 Å		

were checked and included if necessary. Then 25 cycles of maximum-likelihood refinement were performed using REFMAC. The final R_w and R_f values were 17.7 and 20.1%, respectively. All these details are presented in Table 2b.

388 residues. The overall model map correlation coefficient this output was 0.73. A map was calculated using the SOL output phases and 285 peaks were above 3σ cut-off.

Table 2b. Details of PHENIX and ARP/wARP glucose isomerase CuKα data, 11 heavy atom positions. R_w and R_f are in %.

PROGRAM	Resolution limit	10-1.7 Å		
PHENIX	Eleven peaks	Score = 63.0	FOM = 0.29	SOLVE MCC = 0.3626, 557 peaks
	RESOLVE built	341 (335 with side chains)	R_w = 31	R_f = 34 Overall model MCC = 0.71
ARP/wARP		R_w = 30.3	R_f = 30.4	
	No. of auto building cycles	10		
	No. of Refmac cycles in each auto	5		
	building cycle	R_w = 17.5	R_f = 20.9	
	Connectivity Index	0.98		
	No. Chains	3		
	No. Res. Built	382 (with side chains)		
	Water atoms	354		
	R and R_f without water atoms	**R = 23.5**	**R_f = 24.7**	
Final model with solvent atoms		**R_w = 17.7**	**R_f = 20.1**	
		r.m.s. deviation of backbone atoms(1OAD) 0.201 Å (1MNZ) 0.302 Å		

3.2 Glucose isomerase CrKα data

PHENIX was run with the inputs of scalepack format for this data set, single letter sequence file and substructure present in this data as sulphur (S). The imaginary component of the anomalous scattering (f'') of sulphur for this wavelength is 1.14 electron units. The phasing and model building procedure built 383 residues in two chains (256 residues with side chains) out of

To build the side chains of the remaining residues, this o was fed to ARP/wARP and finally it built 381 residues chains out of 388 residues and located 300 water atoms. Ma model building was carried out for the missing residues an water atoms were checked and included if necessary. Th cycles of maximum-likelihood refinement were performed u REFMAC. The final R_w and R_f values were 16.2 and 2 respectively. All these details are presented in Table 3.

Table 3. Details of PHENIX and ARP/wARP glucose isomerase CrKα data, 11 heavy atom po ns. R and R_f are in

	Resolution limit	10-2.3 Å		
	Eleven peaks	Score = 71.8	FOM = 0.37	SOLVE MCC = 0.3452, 285 peaks
	RESOLVE built	383 (256 with side chains)	R_w = 32	R_f = 37 Overall model MCC = 0.73
		R_w = 31.9	R_f = 29.9	
	No. of auto building cycles	10		
	No. of Refmac cycles in each auto	5		
	building cycle	R_w = 16.7	R_f = 21.9	
	Connectivity Index	0.98		
	No. Chains	3		
	No. Res. Built	381 (with side chains)		
	Water atoms	300		
	R and R_f without water atoms	**R_w = 22.5**	**R_f = 25.7**	
Final model with solvent atoms		**R_w = 16.2**	**R_f = 21.3**	
		r.m.s. deviation of backbone atoms(1OAD) 0.204 Å (1MNZ) 0.373 Å		

Results and discussion

Glucose isomerase

Using Co manganese position

Figures 1a and 1b show the cartoon diagrams of the PHENIX output and the final model. The backbone of the final model aligned with superimposed with glucose isomerase structures deposited in the protein data bank (PDB id. 1OAD, space group I 2 2 2; PDB id 1MNZ; space group I222 – atomic resolution one). The root-mean-square deviation is 0.15 Å with PDB (1OAD) and 0.269 Å with PDB (1MNZ). Figure 1c shows a active region of the final model superposed with SOLVE map and final $2|F_o| - |F_c|$ map. The map correlation coefficient between SOLVE map and the final map is 0.299. The average thermal factor of the current model is 14.9 Å^2 and the estimated overall coordinates error is 0.091 Å.

Figure 1a. PHENIX model 191 a.a. (31 a.a. with side chains) from one wavelength CuKα data.

Figure 1b. Final model (CuKα data) using one Mn Auto Built 383 a.a. with side chains using ARP/wARP from PHENIX output.

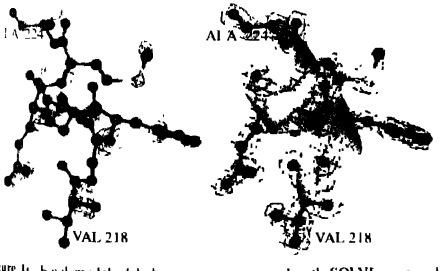


Figure 1c. Final model of helix region superposed with SOLVE map and final $2|F_o| - |F_c|$ map (0.9 σ).

heavy atoms

Figures 2a and 2b show the cartoon diagrams of the PHENIX output and the final model. The r.m.s. deviation of this model with the PDB id 1OAD is 0.201 Å and that with PDB id 1MNZ is 0.302 Å. Figure 2c shows a helix region of the final model superposed with SOLVE map and also final $2|F_o| - |F_c|$ map. The map correlation coefficient between the SOLVE map and the final map is 0.3626. The average thermal factor of the current model is 14.9 Å^2 and the estimated overall coordinates error is 0.096 Å.

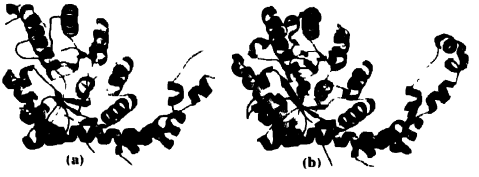


Figure 2a. PHENIX model 341 a.a. (335 a.a. with side chains) from 11 heavy atoms for CuKα data.

Figure 2b. Final model (CuKα data) using 11 heavy atoms Auto Built 382 a.a. with side chains using ARP/wARP from PHENIX output.

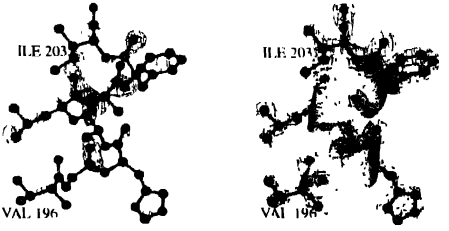


Figure 2c. Final model of helix region superposed with SOLVE map and final $2|F_o| - |F_c|$ map (0.9 σ).

4.2 Glucose isomerase, CrKα data

Figures 3a and 3b show the cartoon diagrams of the PHENIX output and the final model. The r.m.s. deviation of this model

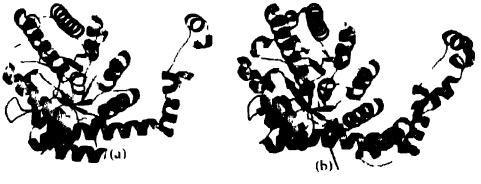


Figure 3a. PHENIX model 383 a.a. (256 a.a. with side chains) from 11 heavy atoms for Cr Kα data.

Figure 3b. Final model (Cr Kα data) using 11 heavy atoms Auto Built 381 a.a. with side chains using ARP/wARP from PHENIX output.

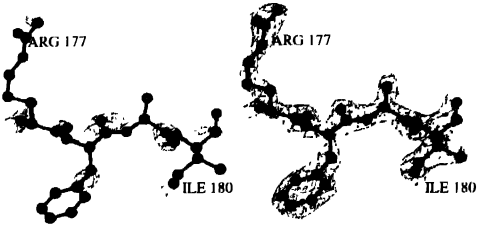


Figure 3c. Final model superposed with SOLVE map and final $2|Fo| - |Fc|$ map (0.7 σ).

with the PDB i d. 1OAD is 0 204 Å and that with PDB i.d. 1MNZ is 0 373 Å Figure 3c shows a section of the final model superposed with SOLVE map and also final $2|F_o| - |F_c|$ map. The map correlation coefficient between the SOLVE map and the final map is 0 3452 The average thermal factor of the current model is 13 8 Å^2 and the estimated overall coordinates error is 0.262 Å

Figure 4a shows the superposition of the Cα atoms of the current model with the final model of CuKα data obtained using one Mn, 11 heavy atoms, final model of CrKα data with PDB i d 1OAD It shows that the overall fold of the models obtained using CuKα and CrKα data sets are similar to that of PDB i.d 1OAD Figure 4b shows the superposition of the current model with the final model using CuKα data set obtained with one Mn, 11 heavy atoms, final model of CrKα data and PDB 1MNZ

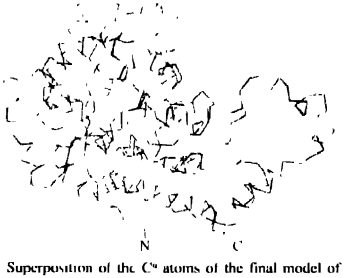


Figure 4a. Superposition of the Cα atoms of the final model of CuKα data using 1Mn (blue), 11 heavy atoms (cyan), CrKα data using 11 heavy atoms (magenta) along with PDB1OAD (red)

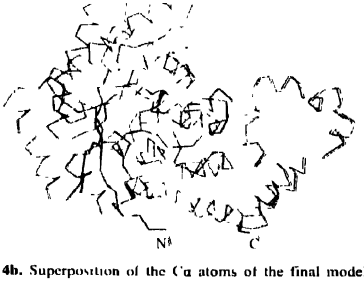


Figure 4b. Superposition of the Cα atoms of the final model of CuKα data using 1Mn (blue), 11 heavy atoms (cyan) CrKα data using 11heavy atoms (magenta) along with PDB1MNZ (red)

5. Conclusion

As suggested by Wang [14], the longer wavelength from a chromium target is advantageous for sulphur SAD phasing because the f'' value of the S atom is larger (1 14 e$^-$) than that obtained with copper radiation (0 57 e$^-$) This can be confirmed from the results obtained for CrKα data, using eleven sulphur positions, PHENIX built only 341 residues for CuKα data whereas 383 residues were built for CrKα data The above work emphasizes the applicability of the SAD techni ie to solve macromolecular structure using lab source dat using CrK radiation when data extends to 2 3 Å resolution SAD dat macromolecule can be used to solve the stru ure with th existing sophisticated program PHENIX in an a omated w using this approach. Automation not only speeds p the pr of solution of crystal structures but also per its fee between various stages of the process, eventu ly l better and more accurate atomic models Inspe tion of P statistics shows that the average quality of structures from the structural genomics centres does not differ quality of structures elucidated in a more traditional way hence this method of solving a macromolecular structure usir lab source data is beneficial to those who do not have access synchrotron data collection.

Acknowledgment

SS thanks Council of Scientific and Industrial Research (CSIR) for providing Senior Research Fellowship DV acknowledge Bioinformatics division of Department of Biotechnology (DB and University Grants Commission (UGC), Govt of India major projects supporting this work and thanks Venture Bus Laboratory authorities, Nagoya Universty, Nagoya, Japan the visiting Professorship assignments in short terms and a acknowledges financial support to the Department under I SAP and DST-FIST programmes

[1] Z Dauter and D A Adamiak *Acta Cryst* **D57** 990 (2001)

[2] U A Ramagopal, M Dauter and Z Dauter *Acta Cryst* D59 (2003)

[3] Z Dauter *Acta Cryst* **D62** 1 (2006)

[4] P D Adams, R W Grosse-Kunstleve, L Hung, T R Ioerger McCoy, N W Moriarty, R J Read, J C Sacchettini, N K Sauter T C Terwilliger *Acta Cryst* **D58** 1948 (2002)

[5] P D Adams, K Gopal, R W Grosse-Kunstleve, L Hung, T R Ic A J McCoy, N W Moriarty, R K Pai, R J Read, T D Rom Sacchettini, N K Sauter, I. C Storoni and T C Terwilli *Synchrotron Rad* **11** 53 (2004)

[6] C S Bond, M P Shaw, S Alphey and W N Hunter *Acta Cry* 755 (2001)

[7] S Selvanayagam, D Velmurugan, T Yamane and A *International Symposium on Recent Trends in Macromolecu Structure and Function Meeting*, **Poster** P24 (2006)

[8] E J Gordon, G A Leonard, S McSweeney and P F Zagalsky *Cryst* **D57** 1230 (2001)

[9] K Sekar, V Rajakannan, D Velmurugan T Yamar Thirumurugan, M Dauter and Z Dauter *Acta Cryst* **D60** 1' (2004)

[10] Z Dauter, M Dauter, E de La Fortelle, G I cogne and G Sheldrick *J Mol Biol* **289** 83 (1999)

[11] Y Kitago, N Watanabe and I Tanaka *Acta Crys* **D61** 1013 (2

[12] A Perrakis, R Morris and V S Lamzin *Nature tructt* (1999)

[13] G N Murshudov, A Lebedev, A A Vagin, K . Wilson and Dodson *Acta Cryst* **D55** 247 (1999)

[14] B –C Wang *Methods Enzymol* **115** 90 (198